Hecht, Martin; Siegle, Thilo; Weirich, Sebastian

A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments

Journal for educational research online 9 (2017) 1, S. 32-51

Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 9 (2017), No. 1, 32–51

Martin Hecht, Thilo Siegle & Sebastian Weirich

A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments

Abstract

Accurate response times are essential for assembling tests in educational large-scale assessment to ensure test validity and efficient testing. Because obtaining empirical response times in pilot studies is cost-intensive and also because this process is complicated in paper-and-pencil assessments, we propose a model-based approach for calculating response times from readily available testlet properties. This prediction formula was developed using the response time data of 334 high school students who worked on 93 testlets of a paper-and-pencil test measuring science achievement. A large proportion (94.3 %) of the variance in response times (i.e., the variation of the average response times of persons across testlets) was explained by number of items, number of words, and response type. Another sample of 1,386 students who worked on 125 additional science testlets was used to validate the initial findings. Overall, the proposed easy-to-use formula is suitable for providing accurate response times for test assembly at a low cost.

Keywords

Response time; Test assembly; Large-scale assessment

Dr. Martin Hecht (corresponding author), Department of Psychology, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
e-mail: martin.hecht@hu-berlin.de

Thilo Siegle, Federal Institute for Educational Research, Innovation and Development of the Austrian School System (BIFIE), Alpenstraße 121, 5020 Salzburg, Austria
e-mail: t.siegle@bifie.at

Dr. Sebastian Weirich, Institute for Educational Quality Improvement (IQB), Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
e-mail: sebastian.weirich@iqb.hu-berlin.de

Ein Modell zur Schätzung von Aufgabenbearbeitungszeiten zur Optimierung von Papier-und-Bleistift-Large-Scale-Assessments

Zusammenfassung

Akkurate Aufgabenbearbeitungszeiten stellen einen essentiellen Bestandteil der Testkonstruktion im Large-Scale-Assessment zur Gewährleistung der Testvalidität und Erhöhung der Effizienz der Testung dar. Da die Erhebung von Aufgabenbearbeitungszeiten in Papier-und-Bleistift-Tests kostenintensiv und kompliziert ist, wird ein modellbasierter Ansatz zur Berechnung der Aufgabenbearbeitungszeiten mittels leicht verfügbarer Aufgabeneigenschaften vorgeschlagen. Diese Vorhersageformel wurde mit Bearbeitungszeitdaten von 334 Schülern, die 93 Aufgaben eines Papier-und-Bleistift-Tests zur Messung von naturwissenschaftlicher Kompetenz bearbeiteten, entwickelt. Ein hoher Anteil (94.3 %) der Varianz der Aufgabenbearbeitungszeiten (d. h., der Variation der mittleren Bearbeitungszeiten von Personen über Aufgaben) wurde durch die Anzahl an Items, Anzahl an Wörtern und dem Aufgabenformat erklärt. Eine zweite Stichprobe von 1386 Schülern bearbeiteten 125 zusätzliche Naturwissenschaftsaufgaben, um die Ergebnisse zu validieren. Insgesamt erwies sich die vorgeschlagene, leicht zu verwendende Formel als geeignet, um akkurate Bearbeitungszeiten zur Testkonstruktion zu geringen Kosten zu liefern.

Schlagwörter

Bearbeitungszeit; Testkonstruktion; Large-Scale-Assessment

1. Introduction

1.1 Theoretical background

Multiple matrix sampling designs are the most commonly applied designs in educational large-scale assessments (Rutkowski, Gonzales, von Davier, & Zhou, 2014). The central idea of such designs is to construct several test forms – called *booklets* in paper-and-pencil tests – that are assembled from a large pool of testlets, which consist of a stimulus and one or several items. A major advantage of this approach is that each individual's workload can be held within acceptable limits while simultaneously covering a variety of different content domains across the test. One essential objective that needs to be fulfilled when compiling booklets is to ensure that the booklet can be reasonably completed within the pre-specified testing time. Therefore, it is pivotal to know the testlet response times that is defined as the average time persons need to complete a testlet. Testlet response time can be obtained in several ways. The most precise testlet response times would obviously be gained from direct measurement in a pilot study. However, this approach

is usually laborious, time-consuming, and costly. Instead, testlet response times are often gauged by didactic experts in the process of testlet construction and development. However, the accuracy of and the consistency between experts' ratings might be – and often is – rather low. A promising alternative is to estimate response times from data that can be accessed without testing, for example, the number of words in a specific testlet. Although extensive amounts of research have addressed a variety of issues concerning response times in educational measurement in recent decades (for comprehensive literature reviews, see Lee & Chen, 2011; Schnipke & Scrams, 2002), surprisingly few studies have broached the idea of obtaining response time estimates from testlet (or item) properties. Halkitis, Jones, and Pradhan (1996) studied the degree to which item response time was related to item difficulty, item discrimination, and word count on a licensing examination. All of the predictors together accounted for half of the variance in the logs of item response time with word count as the strongest predictor (R^2 = 27.2 %), followed by item difficulty (R^2 = 16.2 %), and item discrimination (R^2 = 6.8 %). In the same vein, Bergstrom, Gershon, and Lunz (1994) identified item text length, (relative) item difficulty, item sequence, and position of the correct answer (in multiple-choice items) as relevant predictors. Furthermore, the presence of a figure had a strong impact on response times, although this might have been due to the administration of a separate illustration booklet. In data from a medical licensing examination, approximately 45 % of the variance in item response time was explained by difficulty, the presence/absence of pictures, and the number of words (Swanson, Case, Ripkey, Clauser, & Holtman, 2001). The authors reported that "a logit change in item difficulty adds 14+ seconds", "the presence of a picture adds 12+ seconds," and "each word adds approximately 0.5 seconds" (p. 116). Even though empirical studies on this topic are rare, the results indicate that predicting response times from item properties is a worthwhile endeavor.

Test construction is not an end in and of itself but is always conducted with the goal of testing a specific population of students. Here, response times can provide valuable information about how to design the test as tests may function differently in different subpopulations. Consequently, this information is useful for tailoring tests to fit the needs of subpopulations with different time requirements. Research on the relations between person properties and response times is much more elaborate than research on item properties (again, see Lee & Chen, 2011; Schnipke & Scrams, 2002). However, the question of how student characteristics influence response times is typically addressed from a different angle with research that treats response time estimates as an auxiliary source of information for estimating individual ability (e.g., Wang & Hanson, 2005). Conversely, a person's ability is particularly important when studying response times. In a pioneering article on the estimation of response times (Thissen, 1983), the ability-latency relation was strongly moderated by the test content. Correlations between effective ability and slowness ranged from zero for a spatial visualization test to .94 for a figural reasoning task. Analyses with contemporary statistical models (e.g., Klein Entink, Fox, & van der Linden, 2009) confirmed the complexity of this connection: Some studies found a

negative relation between ability and speed, indicating that more capable test-takers spent more time on a task (Goldhammer & Klein Entink, 2011), whereas others reported the opposite result (Davison, Semmes, Huang, & Close, 2011). Besides the speededness of the measure, the relation has been further moderated by the anticipated consequences (low- vs. high-stakes testing) and personality traits such as conscientiousness or impulsivity (also see research on the *speed-accuracy tradeoff*, e.g., Goldhammer, 2015 for an overview). Furthermore, persons differ in mental speed of information processing – this individual baseline speed is an important factor when investigating response times and should be taken into account to avoid bias (Mayerl, 2005). In summary, the relations between response times and student properties are not clear.

The popularity of research on response times has soared with the advent of the technology to measure them directly in computer-based assessments. Obviously, measuring response times in paper-and-pencil settings is far more complicated, and this is presumably the reason that almost all studies rely on computer-based data. However, data from computer-based tests may not be suitable for assembling paper-and-pencil tests because transposing the content from computer to paper may affect the reliability and validity of the measure. Although meta-analyses on the comparability of paper-based and computer-based assessments have reported only small to negligible cross-mode differences (e.g., Mead & Drasgow, 1993; Wang, Jiao, Young, Brooks, & Olson, 2007, 2008), three caveats must still be considered when interpreting such findings. First, this comparability holds only for unspeeded measures as Mead and Drasgow (1993) conclusively demonstrated that the almost perfect cross-mode correlation for timed power tests dropped considerably to .72 for speeded tests. Second, meta-analyses usually consider the mean structure but not the variance-covariance structure. Even if there are no mode effects for means, there might be mode effects concerning the variances and covariances (Schroeders & Wilhelm, 2011). Third, whereas cross-media differences are small in general, in a specific instantiation, substantial differences between test media may occur (van Lent, 2008), and without generalizable knowledge about which factors affect the equivalence, it is difficult to determine the impact that a transition will have on response times. Factors affecting response times across media might consist of differences in the perceptual demands or the motor-skill requirement in the response procedure (Schroeders & Wilhelm, 2010). More precisely, differences across administration modes can result from scrolling down long texts on small screens with low screen resolution (Bridgeman, Lennon, & Jackenthal, 2003), clicking response buttons with a mouse instead of ticking the solution on a sheet of paper with a pen (Pomplun, Frey, & Becker, 2002), and using a keyboard instead of answering manually (Overton, Taylor, Zickar, & Harms, 1996). In summary, the change from paper to computer may alter the construct that the test administrator intends to measure. With this concern in mind, we decided not to assess response times on computers but to employ a paper-and-pencil assessment instead.

1.2 The scope of the present research

The aim of the present study was to provide a well-founded and easy-to-use formula to calculate response times for testlets, stimuli, and items in order to optimize the assembly of paper-and-pencil tests in educational large-scale assessments. Furthermore, we explored whether response times depended on certain person properties in order to determine how to tailor test construction to the specific needs of specific subgroups of students. More precisely, we modeled response times as dependent on the testlet properties (a) number of items, (b) number of words, (c) response type (multiple choice, short response, extended response), and (d) testlet difficulty. We simultaneously regressed them on the following student properties as well: (a) sex, (b) school track, and (c) competence. Furthermore, two-way interactions of student and testlet properties were investigated exploratively. In a second step, we validated this empirically obtained model in a new sample of testlets and students.

2. Method

2.1 Participants

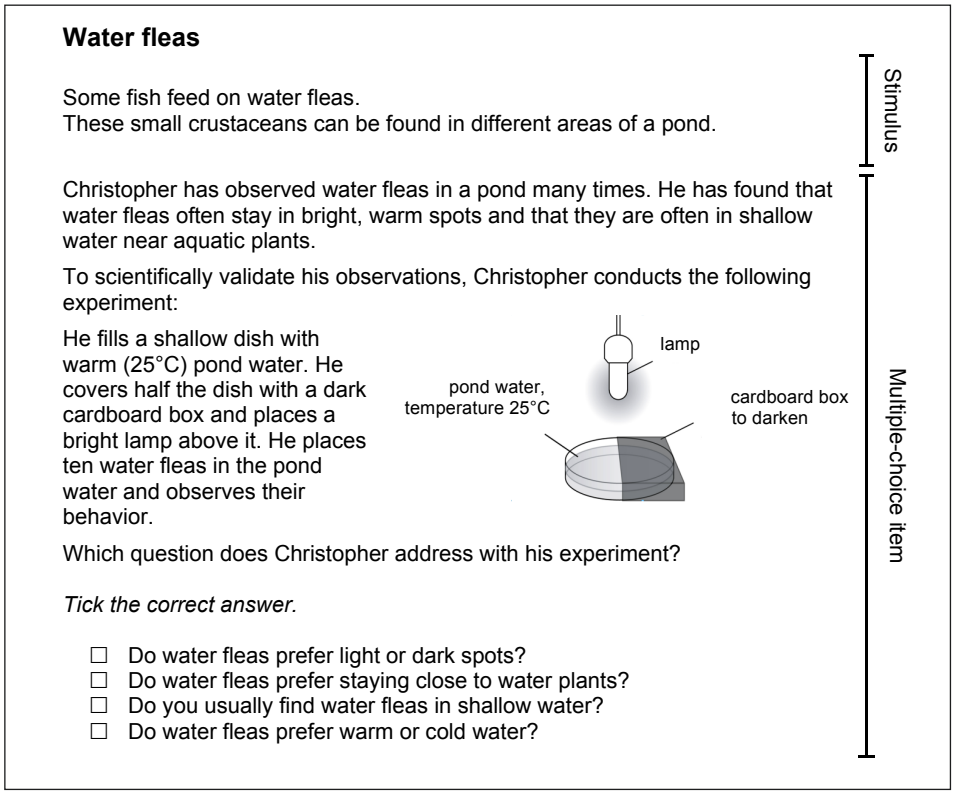
Study 1 was used to develop the prediction formula. The sample consisted of 334 students in Grade 9 (49.4 % girls, 2.1 % did not indicate their sex) with an average age of 15.5 years (SD = 0.75) from four academic-track schools (56.9 %) and three intermediate-track schools (43.1 %). Academic-track schools prepare students for university enrollment, whereas students in intermediate-track schools often pursue a vocational education. Participation was voluntary, and students were not rewarded or graded in any way. Data were collected in the spring of 2010.

Study 2 was conducted for validation purposes. The data collection took place in the fall of 2010. All 1,386 students were 10th graders from intermediate-track schools, and almost half of them were girls (48.0 %, 1.9 % did not indicate their sex).

2.2 Design and procedure

In Study 1, we distributed 93 testlets, each of which contained a stimulus and one to five (M = 1.86) items. They originated from a large pool of testlets that were designed to measure the *German Science Education Standards* (for details on the development and evaluation of these standards, see Kremer et al., 2012; Neumann, Fischer, & Kauertz, 2010; Pant et al., 2013). The conceptual core of educational standards is very similar to the idea of *scientific literacy* (e.g., Holbrook & Rannikmae, 2009) and contains four subdomains: content knowledge, scientif-

Figure 1: Example of a science testlet consisting of a stimulus and a multiple-choice item



ic inquiry, decision making, and communication. Because the test development for these subdomains was time-delayed, only testlets measuring content knowledge and scientific inquiry were available in Study 1. The required response types consisted of either choosing an answer (multiple choice), writing one or several words (short response), or writing several sentences (extended response). Figure 1 displays an example testlet from the subdomain *scientific inquiry* consisting of a stimulus and a multiple-choice item. We employed an incomplete block design (e.g., Frey, Hartig, & Rupp, 2009) with 24 booklets that were randomly administered to the students. The test construction process began by grouping testlets into clusters of 20 min. Eight clusters were assembled for each science subject (biology, chemistry, physics). In the next step, each of these clusters was assigned to two booklets. Following this procedure, each booklet contained three clusters, one for each science domain. Because the unspeeded response time was the variable of interest, all students were provided with sufficient time to complete the test. Before and after working on each testlet, students were asked to record the time in the test booklet. In order to standardize the time recording, a clock was positioned in front of the class. Leaving the time recording to students worked surprisingly well. Only a very few data points had to be removed due to unreadability or implausibility.

In Study 2, the design and procedure were similar to Study 1, except for three changes: (a) the topic of the testlets consisted of another scientific competence: decision making, (b) booklets contained two clusters of 20 min length equaling 40 min of total testing time, and (c) booklets contained either one cluster of chemistry and one cluster of physics or two clusters of biology. A total of 51 booklets were assembled and randomly administered to the students.

2.3 Statistical analyses

Response times are on the Person × Testlet level. As a consequence of the employed multiple matrix sampling design each person responded to several, but not all testlets. Thus, persons and items are partially crossed (see, for instance, Hecht, Weirich, Siegle, & Frey, 2015a). A suitable data analysis technique for crossed data structures is *linear mixed models*[1] (LMM). We specified five consecutive LMMs to predict the response time using the characteristics of testlets and students. Response times were recorded in seconds – y_{jt} was the time student j worked on testlet t. Because booklets were distributed to students randomly missingness due to the multiple matrix design was *completely at random* (MCAR). LMM software such as the R (R Core Team, 2014) package *lme4* (Bates, Mächler, Bolker, & Walker, 2014) is able to handle MCAR adequately.

The first model in the series contained only an intercept α_o, a student parameter, θ_j, a testlet parameter, β_t, and a Student × Testlet interaction parameter, ε_{jt}:

$$y_{jt} = \alpha_0 + \theta_j + \beta_t + \varepsilon_{jt} \quad (1)$$

In this model, the intercept α_o is the overall mean testlet response time, θ_j is the deviation of student j from this mean, and β_t is the deviation of testlet t. The term ε_{jt} is the interaction of a specific student j with a specific testlet t. As the purpose of the present study was to investigate the effects of testlet and person properties on the response time, the point estimates for students and testlets were of less interest. Thus, students, testlets, and interactions were each modeled as *random effects*, assuming a normal distribution with means of zero and variances of σ^2_θ, σ^2_β, and σ^2_ε. These variances indicated the extent to which the students and testlets diverged from the overall mean on average. Taking this individual baseline speed into account is pivotal for unbiased analyses (Mayerl, 2005). Further, testlet and student properties in subsequent models were expected to explain this variability in response time.

In the second model, the number of items ($N_{t,items}$) and the total number of words ($N_{t,words}$) in a testlet t were added to the model as predictors (*fixed effects*):

1 Linear mixed models (e.g., McCulloch, Searle, & Neuhaus, 2008) are related to multilevel (or hierarchical) models (e.g., Raudenbush & Bryk, 2002). Multilevel models are special linear mixed models and thus can be described within the linear mixed model framework.

$$y_{jt} = \alpha_0 + \gamma_{items} N_{t,items} + \gamma_{words} N_{t,words} + \theta_j + \beta_t + \varepsilon_{jt} \quad (2)$$

where the intercept α_0 is the mean response time of the (hypothetical) testlets with zero items and zero words. The variance of testlets, σ^2_β, is conditional on the effects of testlet properties and could be interpreted as the remaining unexplained variance. The main purpose of Model 2 was to estimate the additional time that a single item added to the response time needed to complete a testlet. The estimate of this effect, γ_{items}, was used as a *fixated* effect in the following models to facilitate interpretation and to avoid estimation problems due to matrix rank deficiency. Note that the term *fixated* and the asterisk sign * is used for fixed effects that are fixed to a specific value. Hence, γ^*_{items} in the next models is the estimated value of γ_{items} from Model 2.

Model 3 additionally included the numbers of multiple-choice items ($N_{t,MC}$), short response items ($N_{t,SR}$), and extended response items ($N_{t,ER}$) and the centered difficulty of the testlet ($X_{t,diff}$) as predictors:

$$y_{jt} = \alpha_0 + \gamma^*_{items} N_{t,items} + \gamma_{words} N_{t,words} + \gamma_{MC} N_{t,MC} + \gamma_{SR} N_{t,SR} + \gamma_{ER} N_{t,ER} + \gamma_{diff} X_{t,diff} + \theta_j + \beta_t + \varepsilon_{jt} \quad (3)$$

Whereas the effect of number of items indicates how much response time is needed for an item in general, the response type effects express the additional time needed for items of a particular response type. The testlet difficulty (i.e., the average of the threshold parameters), $X_{t,diff}$, was estimated using a partial credit model with the software package ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007) and centered afterwards. Thus, if the difficulty changed by one logit, the response time increased by a value of γ_{diff}.

In Model 4, student properties were added to explain the variability in students' response times. The predictors were sex (Z_{sex}), school track (Z_{track}), and students' competence (Z_{comp}). We used the Greek letter δ instead of γ to distinguish between the effects of student properties and testlet properties:

$$y_{jt} = \alpha_0 + \delta_{sex} Z_{sex} + \delta_{track} Z_{track} + \delta_{comp} Z_{comp} + \gamma^*_{items} N_{t,items} + \gamma_{words} N_{t,words} + \gamma_{MC} N_{t,MC} + \gamma_{SR} N_{t,SR} + \gamma_{ER} N_{t,ER} + \gamma_{diff} X_{t,diff} + \theta_j + \beta_t + \varepsilon_{jt} \quad (4)$$

The variability in students was modeled as conditional on the effects of student properties and represented the unexplained variability in students' response times. Students' competence estimates (centered WLEs, i.e., weighted likelihood estimates; Warm, 1989) came from the same item response model as the testlet difficulties. The dichotomous variables sex (boys vs. girls) and school track (intermediate vs. academic) were effect coded. Instead of using the default effect codes (i.e., -1 and 1), we modified them according to the proportions of the respective groups

in our sample (this is essentially equivalent to centering the effect codes). For the variable school track, the centered effect codes were Z_{track1} = -1.14 for the intermediate track and Z_{track2} = 0.86 for the academic track. Because the sample contained almost equal proportions of boys and girls, the centered effect codes for the variable sex were Z_{sex1} = -1.01 for boys and Z_{sex2} = 0.99 for girls. The advantage of centered effect codes is that the effect is estimated for equal proportions of the two groups (50 %) even though the distributions in the sample may be different. As a consequence, the intercept does not change when such centered effect coded variables are entered into the model. Furthermore, we included all significant Student × Testlet interactions in the model by first including all interactions and then dropping insignificant ones. For the sake of clarity, Equation 4 does not contain the interaction terms and in Table 3, results are only reported for interactions that have been found significant.

For the final model, Model 5, all nonsubstantial effects from Model 4 were excluded to derive a prediction formula that could be easily implemented to calculate the response times for testlets and items on paper-and-pencil tests in large-scale assessments.

All models were estimated with the function *lmer* from the *lme4* package (Bates et al., 2014). Confidence intervals were bootstrapped using the lme4 function *bootMer* with 10,000 simulations and the function *boot.ci* from the package *boot* (Canty & Ripley, 2014). An estimate is considered significantly different from zero if zero is outside the 95 % confidence interval. The main reason for using the bootstrapped and therefore potentially asymmetric confidence intervals is that *lme4* does not provide standard errors for variance estimates because "in most cases summarizing the precision of a variance component estimate by giving an approximate standard error is woefully inadequate" (Bates, 2010, p. 19).

The log of the response times is often reported as log-normally distributed and analyzed accordingly. For the data at hand, the log of the response times did not better approximate a normal distribution. Thus, the original response time estimates were used.

3. Results

3.1 Descriptive statistics

Table 1 shows the descriptive statistics for the empirical testlet response time and testlet properties in Study 1 (for prediction) and Study 2 (for validation). The main difference between the testlets used in these two studies was their length. The Study 2 testlets, which measured decision-making competence in science, contained more items (M_1 = 1.86 vs. M_2 = 2.92) and nearly twice as many words (M_1 = 175.14 vs. M_2 = 322.82) than the Study 1 testlets, which measured content knowledge and scientific inquiry skills. These differences in testlet length were

reflected in a much higher average response time in Study 2 (M_1 = 148.20 s vs. M_2 = 277.66 s). The correlations between the empirical testlet times and testlet properties are shown in Table 2. Not surprisingly, the number of items was highly correlated with response time (r_1 = .87, r_2 = .70). The correlations between the number of words and response time were also large (r_1 = .75, r_2 = .64). Further, the number of items and the number of words were also substantially correlated (r_1 = .68, r_2 = .65). The correlations between the response type variables and the number of words ranged from .32 to .50 in Study 1.

Table 1: Descriptive statistics for testlet response time and testlet properties in Studies 1 and 2

Testlet characteristic	Study 1 (N = 93)				Study 2 (N = 125)			
	M	*SD*	Min.	Max.	*M*	*SD*	Min.	Max.
Testlet response time	148.20	66.68	57.07	320.00	277.66	91.03	81.43	621.00
Number of items	1.86	1.04	1	5	2.92	1.32	1	7
Number of words	175.14	82.43	27	450	322.82	123.39	106	733
Multiple-choice items	1.12	0.88	0	4	1.03	1.05	0	5
Short response items	0.47	0.75	0	3	0.78	0.91	0	4
Extended response items	0.27	0.53	0	2	0.98	1.18	0	5
Testlet difficulty	-0.14	1.01	-3.09	4.85	0.47	1.34	-2.33	4.21

Notes. Testlet response time is presented in s. Testlet difficulty is presented in logits.

Because of the high intercorrelations between the number of items and the other variables, the effect was first estimated in Model 2 and then fixated in all subsequent models. This approach allowed us to disentangle the impact of number of items and the other variables on response time despite the high correlations. To provide a better understanding of the relations between variables, partial correlations (i.e., correlations controlled for the number of items) are reported in the upper triangle of Table 2 (calculated with the R package *parcor*; Krämer & Schäfer, 2014). For example, as the number of multiple-choice items increased – relative to items with other response types – the testlet response time decreased in both studies (r_1 = -.43, r_2 = -.58). By contrast, as the number of extended response items increased, the testlet response time was higher (r_1 = .27, r_2 = .48). The relation between testlet difficulty and response type was as follows: Testlets that contained more multiple-choice items were easier (r_1 = -.24, r_2 = -.41), whereas those that contained more extended response items were associated with greater testlet difficulty (r_1 = .35, r_2 = .45).

Table 2: Correlations of testlet response time and testlet properties in Studies 1 and 2

Testlet characteristic	Study	Testlet characteristic						
		1	2	3	4	5	6	7
Testlet response time (1)	1			.42	-.43	.23	.27	.27
	2			.33	-.58	.00	.48	.47
Number of items (2)	1	.87						
	2	.70						
Number of words (3)	1	.75	.68		.38	-.26	-.20	.00
	2	.64	.65		.14	.00	-.13	.09
Multiple-choice items (4)	1	.32	.56	.62		-.74	-.45	-.24
	2	-.13	.36	.34		-.19	-.63	-.41
Short response items (5)	1	.50	.45	.13	-.30		-.23	.00
	2	.23	.30	.20	-.06		-.60	-.11
Extended response items (6)	1	.48	.38	.13	-.14	-.02		.35
	2	.62	.41	.17	-.39	-.40		.45
Testlet difficulty (7)	1	.11	-.05	.02	-.24	-.01	.32	
	2	.36	.04	.13	-.38	-.10	.43	

Notes. Correlations are displayed in the lower triangle. Partial correlations with the number of items partialed out are displayed in the upper triangle.

3.2 Models

Table 3 displays the results of all five consecutive models. In Model 1, the intercept representing the overall average testlet response time was α_o = 149.4 s. The deviations of testlets ($\bar{\sigma}_\beta$ = 65.1, 95 % CI [55.4, 75.4]) and students ($\bar{\sigma}_\theta$ = 27.4, 95 % CI [24.0, 30.9]) were significantly different from zero – thus, there was indeed a substantially large amount of variability that could be explained by the properties of testlets and students in further models. The purpose of Model 2 was to estimate the effect of the number of items so that the parameter could be included as a fixated effect in subsequent analyses. This effect amounted to γ_{items} = 43.8 s, which means that adding one item to the testlet increased the response time by 43.8 s on average. The two testlet properties in Model 2, number of items and number of words, explained 83.4 % of the variability in testlet response time. Further, Model 2 possessed a smaller Bayesian information criterion (BIC) and Akaike information criterion (AIC) than Model 1 (see Table 3), indicating that this model was more suitable for explaining the data at hand. In Model 3, the number of items of any response type (i.e., multiple choice, short response, or extended response) and the centered testlet difficulty were added as predictors. The effect of the *multiple-choice* response type was estimated as γ_{MC} = -24.4 s, that is, students were able to answer multiple-choice items faster than the overall average. In other

words, adding one multiple-choice item increased the response time by $\gamma_{items} + \gamma_{MC}$ = 43.8 - 24.4 = 19.4 s. The effects for short response and extended response items were γ_{SR} = 2.8 and γ_{ER} = 14.5, respectively. Comparing a multiple-choice item to an extended response item (with an equal number of words and difficulty) yielded a difference of 14.5 - (-24.4) = 38.9 s because writing a paragraph takes much longer than just ticking boxes. Surprisingly, the difficulty of the task played no role as indicated by the near-zero and nonsignificant effect γ_{diff} = 1.1, 95 % CI [-3.3, 5.4]. The intercept was also near zero (α_o = 2.6, 95 % CI [-11.4, 16.2]) because a hypothetical testlet with zero words and zero items would take no time to complete. Further, a (hypothetical) testlet with just a single stimulus but no items had a response time that depended on only the words of the stimulus that needed to be read. For every word in the testlet, the response time increased by γ_{words} = 0.37 s. Thus, increasing the text length by 100 words would add 37 s to the predicted response time.

Besides the testlet properties, Model 4 additionally included student properties. All student properties that were considered in our study – sex, school track, and students' competence – exhibited only very marginal nonsignificant effects (δ_{sex} = 0.91, 95 % CI [-3.0, 4.7], δ_{track} = 1.6, 95 % CI [-2.5, 5.8], and δ_{comp} = 0.18, 95 % CI [-4.4, 4.7]). Therefore, it did not seem necessary to adjust the response times for booklets that were specifically designed for these subsamples (boys vs. girls, intermediate vs. academic track, more vs. less competent students). However, two interactions between testlet and person properties were relevant and therefore included in Model 4, that is, Sex × Extended Response (6.1, 95 % CI [1.6, 10.6]) and School Track × Extended Response (16.6, 95 % CI [12.2, 21.1]). Thus, girls worked 2 * 6.1 = 12.2 s longer on extended response items than boys and students who were enrolled in an academic-track school worked 2 * 16.6 = 33.2 s longer on an item with an extended response format than students enrolled in an intermediate-track school. These differences should be taken into account when tests contain a sufficient number of extended response items and need to be tailored to these subgroups.

For Model 5, all of the nonsubstantial predictors from Model 4 were excluded to derive a formula that would be easy to use to calculate response time estimates. Although the effect of short responses was nonsignificant, it was retained in the prediction model in order to facilitate confusion-free handling. This final and best fitting model (lowest AIC and BIC) explained 94.3 % of the variability in testlets:

$$\hat{y}_t = 43.8N_{t,items} + 0.39N_{t,words} - 25.9N_{t,MC} + 2.2N_{t,SR} + 14.7N_{t,ER} + 6.1Z_{sex}N_{t,ER} + 16.6Z_{track}N_{t,ER} \quad (5)$$

3.3 Examples

We will now present examples that show how to use this formula, which can be applied to calculate response times for (a) stimuli, (b) items, and (c) testlets. In the simplest case of a stimulus, all predictors are set to zero except for

Table 3: Fixed and random effect estimates, explained variance, and model fit of linear mixed models

	Model 1		Model 2		Model 3		Model 4		Model 5	
	Est.	95 % CI	Est.	95 % CI	Est.	95 % CI	Est.	95 % CI	Est.	95 % CI
Fixed effects										
Intercept	149.4	[135.6, 163.1]	28.2	[13.8, 42.7]	2.6	[-11.4, 16.2]	2.4	[-11.6, 16.0]		
Number of items			43.8	[36.0, 51.7]	43.8*	—	43.8*	—	43.8*	—
Number of words			0.23	[0.13, 0.32]	0.37	[0.30, 0.45]	0.37	[0.30, 0.45]	0.39	[0.36, 0.41]
Multiple-choice items					-24.4	[-32.0, -17.1]	-24.7	[-32.2, -17.4]	-25.9	[-31.2, -20.7]
Short response items					2.8	[-3.5, 9.1]	2.8	[-3.5, 9.2]	2.2	[-3.4, 8.0]
Extended response items					14.5	[6.6, 22.9]	14.6	[6.6, 22.9]	14.7	[7.3, 22.4]
Testlet difficulty					1.1	[-3.3, 5.4]	1.0	[-3.4, 5.4]		
Sex (girls)							0.91	[-3.0, 4.7]		
School track (academic)							1.6	[-2.5, 5.8]		
Students‘ competence							0.18	[-4.4, 4.7]		
Sex × Extended Resp.							6.1	[1.6, 10.6]	6.1	[1.6, 10.6]
School Track × Extended Resp.							16.6	[12.2, 21.1]	16.6	[12.2, 21.1]
Random effects										
Testlets	65.1	[55.4, 75.4]	26.5	[22.3, 31.7]	15.4	[12.6, 19.9]	15.4	[12.7, 20.0]	15.5	[12.5, 19.7]
Students	27.4	[24.0, 30.9]	27.4	[24.0, 30.9]	27.4	[24.1, 31.0]	27.4	[24.3, 31.2]	27.4	[24.2, 31.0]
Students × Testlets	71.4	[69.6, 73.2]	71.4	[69.6, 73.2]	71.4	[69.6, 73.2]	70.7	[69.0, 72.5]	70.7	[69.0, 72.5]
Explained Variance										
Testlets			83.4 %		94.4 %		94.4 %		94.3 %	
Students			0.0 %		0.0 %		0.0 %		0.0 %	
Students × Testlets			0.0 %		0.0 %		1.9 %		1.9 %	
Model fit										
AIC (diff1, diff2)	41363	—	41215	(-149, -149)	41146	(-69, -218)	41092	(-54, -271)	41083	(-9, -280)
BIC (diff1, diff2)	41388	—	41252	(-136, -136)	41202	(-50, -187)	41179	(-23, -209)	41139	(-40, -249)
Deviance (diff1, diff2)	41355	—	41203	(-153, -153)	41128	(-75, -228)	41064	(-64, -291)	41065	(1, -290)

Notes. CI = confidence interval; diff1 = difference of model fit index in reference to the previous model; diff2 = difference of model fit index in reference to Model 1. Fixated effects are tagged with *. For random effects, the estimate reported is the *SD*.

the number of words. The response time for a stimulus with $N_{t,words}$ = 100 is then $y_{(a)}$ = 0.39 * 100 = 39 s. An extended response item with $N_{t,words}$ = 100 will take $y_{(b)}$ = 43.8 * 1 + 0.39 * 100 + 14.7 * 1 = 97.5 s to complete. If this item will be employed in academic-track schools, 14.3 s should be added: $y_{(b)2} = y_{(b)} + 16.6 * Z_{track2} * N_{t,ER}$ = 97.5 + 16.6 * 0.86 * 1 = 111.8 s. For intermediate-track schools, 18.9 s should be subtracted: $y_{(b)1} = y_{(b)} + 16.6 * Z_{track1} * N_{t,ER}$ = 97.5 + 16.6 * (-1.14) * 1 = 78.6 s. For a testlet, two approaches are feasible: either applying the formula to the entire testlet or summing the response times of the elements. We combined the previously used stimulus and two of the previously used extended response items into a testlet. When applying the formula, this yielded: $y_{(c)1}$ = 43.8 * 2 + 0.39 * (100 + 100 + 100) + 14.7 * 2 = 234 s. Alternatively, the separately calculated response times can be summed across the three elements: $y_{(c)2} = y_{(a)} + 2 * y_{(b)}$ = 39 + 2 * 97.5 = 234 s. An asset of this formula is that it allows various testlets to be assembled from items with pre-calculated response time without the need to apply the formula to the testlet.

3.4 Validation

To validate the empirically derived prediction formula, another sample of persons completing another sample of N = 125 testlets were drawn and testlet response times recorded. Thus, empirical testlet response times can be compared to predicted testlet response times that were calculated by plugging testlet properties into Equation 5. For each of the 125 testlets the difference between the predicted and empirical response time (prediction bias) was calculated. The mean prediction bias (across testlets) was M_{bias} = -34.52 (SD_{bias} = 49.88). Thus, response times are underestimated by M_{bias} / $M_{empirical}$ = -34.52 / 277.66 = -12.4 %. The root mean square prediction error, which takes both the bias and the variation of predicted values into account, is RMSPE = 60.50.

4. Discussion

Accurate response times of testlets and items are crucial for assembling the booklets of a paper-and-pencil test that will be used in large-scale assessments. The most accurate response times can certainly be obtained by pilot testing the testlets and items, but this process is time-consuming and expensive. Nevertheless, even for pilot testing, it is necessary to have some initial response time estimates for testlets. Obtaining response time estimates from available testlet and item properties is a quick, convenient, and low-cost alternative to extensive pilot testing or expert ratings. To derive an empirically based formula, we collected response times from a sample of high school students who worked on science testlets. On the basis of these empirical data, we acquired a sound prediction model (Model 5) that

can be used to estimate response times for stimuli, items, and testlets from (a) the number of items, (b) the number of words, and (c) the response type. These are all easy-to-obtain properties that are available without cost-intensive pilot testing.

Our results are plausible and in line with previous research. Number of words was also identified as a relevant predictor in the studies by Halkitis et al. (1996), Bergstrom et al. (1994), and Swanson et al. (2001). In our data, it took 0.39 s to process one word, a value that is close to Swanson's estimate of (approximately) 0.5 s. In our assessment of student properties, we found no main effects of sex or school track. These findings are consistent with Bergstrom et al.'s (1994) findings which suggested that "examinee characteristics are generally not related to response time" (p. 13). However, in light of the person variation of $\widehat{\sigma}_{\theta} = 27.4$ in our model, the statement that person characteristics are *generally* not related to response times is rather questionable. In our data, there are differences between persons, and they would probably be explainable if one only had the "right" predictors. Nonetheless, for test designers, the null effect of sex and school track is a satisfactory outcome because there is no need to construct separate test forms, for example, for academic-track and intermediate-track schools. A somewhat puzzling result in our study was the null effect of testlet difficulty, whereas in other studies this was a weak to moderately strong predictor. There might be several reasons for this. First, difficulty and response time were assessed on the level of testlets (instead items). Thus, there might be more expectable effects on the item level that cancel themselves out when aggregated to testlets. Second, there might be individually varying effects of testlet difficulty that sum to zero on average. For example, more competent students might be faster in answering difficult items. However, if less competent students tend to quickly skip a testlet they are then again as fast as (or even faster than) more competent students. Such a phenomenon might also explain the null effect of the Testlet Difficulty × Student's Competence interaction in our study. Of course, there might be more person characteristics that moderate the difficulty effect. Future research should explore Testlet/Item × Person interactions more thoroughly. Further, the null effect of testlet difficulty in this study does not imply that testlet difficulty is a negligible variable when assembling large-scale tests. Although students will approximately need the same time independent of testlet difficulty, compiling too difficult or too easy booklets might result in problematic booklet effects (Hecht, Weirich, Siegle, & Frey, 2015b).

The statistical method (linear mixed models) that we employed allowed us to estimate Item × Student interactions. We investigated these in an exploratory fashion and found that girls and academic-track students invested more time in writing extended responses. Test designers selecting test items should be aware of such additional influences on response time. Booklets with a disproportionately high number of such items may differentially affect the response times of students in specific educational tracks. However, we would like to suggest that test administrators carefully consider the political implications of allocating different times to subgroups because questions of test fairness may emerge. Furthermore, student estimates from studies with different time restrictions may be difficult to compare.

However, using the final prediction formula to assemble and optimize test booklets that are administered to all students from a certain population should not be problematic.

To validate our empirically derived prediction formula, a second sample of students worked on a different set of science testlets. Applying Equation 5 and comparing predicted and empirical response times yielded an average underestimation of -12.4 %. This prediction bias might have several reasons as Study 1 and Study 2 differed in certain aspects, for instance, students' grade (9 vs. 10) and test length (60 min vs. 40 min). Further, different competences were measured in the two studies. Whereas testlets that measured content knowledge and scientific inquiry were used in the original sample, in the validation study (Study 2), the competence in question was decision making in science. Such items require test-takers to thoroughly elaborate on a decision or an evaluation, a process that appears to be more time-consuming than answering the Study 1 items, which assessed knowledge about science and scientific procedures. An inspection of students' responses to the extended response items suggested that students indeed wrote more when they answered items from the domain decision making. Such a potential effect of the content domain is undeniably one of the major threats to the chosen prediction model because the response type extended response is just a rough proxy for the actual amount of text that is produced. Depending on the competence that is measured or on other (unknown) variables, there might be nontrivial variation in the amount of text students produce. In other words, generalizability to other competence domains might be limited and users should exercise caution when applying the prediction formula to very different content domains. However, as our validation study showed, within a relatively wide range of competencies (content knowledge, scientific inquiry, and decision making) the formula works quite well. Of course, it is not applicable if items are not in either multiple choice, short response, or extended response format.

Moreover, the results of the present study may not generalize to populations other than German-speaking students or students in Grade 9. Sentences in other languages might be either more concise or lengthier and thus faster or slower to read and write. Furthermore, younger (e.g., primary school) students might be much slower at reading the same amount of text. A further limitation was the use of response times from paper-and-pencil assessments because such measurements are less precise in comparison with a computer-based assessment. On the other hand, using computer-based response times to construct paper-and-pencil tests is also not a feasible option as mode effects may jeopardize the validity of the measurement and lead to severe biases. More research is needed to investigate and predict mode effects on response times and to describe their implications. Further, potentially less precise paper-and-pencil response times would just add "noise" to the relations under investigation. Thus, relations would appear smaller than they actually were if response time measurement was accurate. Given the large amount of explained variance (94.3 %), it is reasonable to assume that our measurements

were quite accurate. Still, the reported results are lower boundaries and may even be more pronounced in other studies with even higher measurement precision.

In this study, we used a modeling approach that explicitly allowed us to divide the response time variance into variance accounted for by items and variance accounted for by persons. The presented mixed models offer three methodological advantages over the often-used standard regression analyses. First, mixed models enabled us to consider Testlet × Student interactions that could provide additional information for test construction. Second, mixed models can adequately account for the data structure in large-scale assessments where students work on different item subsets assembled in various booklets. Third, mixed models offered higher test power because the response time data were not aggregated (and thus not reduced) across persons or items. In other words, more data points were available for the estimation of model parameters. A limitation of our models is the assumption of equal testlet variances (homoscedasticity). Within our modeling framework, this assumption could have been easily tested. We did not pursue this, because for the purpose of predicting mean testlet response time it is – or should be – rather irrelevant. Modeling heteroscedasticity would have merely improved model fit, but fixed effects parameters would have been approximately the same. Furthermore, in our models, we fixated the effect of number of items to allow for a certain interpretation of effects (i.e., deviation of a certain response format from the mean time of a testlet with a certain number of items). Other models are imaginable where number of items is excluded. This would just change the interpretation of the response format effects, but would otherwise result in similar effects after converting them accordingly.

Another important methodological issue was the criterion that was targeted in our prediction model. In line with other studies, we predicted the mean response time. This implies that 50 % of the students will complete the testlet in this amount of time, and 50 % will not (under the assumption of a normal distribution). Test administrators should consider whether the mean response time is the correct choice for the specific application of the test. One may argue that some other criterion will offer a worthwhile alternative. For instance, it may be reasonable to enable 90 % of the students to complete the testlet, in which case the .90 quantile of the response time distribution would be chosen. A related – and rarely discussed – issue is the aggregation of item or testlet response time into booklet response time. Here, the standard approach is to sum the response times across items or testlets to derive the booklet response time, which equals the time that is available per student. This technique might not work for all criteria because students' rank ordering of response times will change from testlet to testlet if their correlations are below 1 (which is usually the case). This implies that testlet time cannot simply be added to calculate booklet time if certain criteria are used. For instance, if 90 % of the students are expected to complete their booklets, it is not correct to sum the .90 quantiles of the testlets that comprise the booklet. Instead, some lower quantile would be the right choice in this case. Further research is needed to identify the testlet quantiles that lead to a certain target quantile at the booklet level.

To conclude, the present study provides an empirically derived formula for the prediction of response times for items, stimuli, and testlets for paper-and-pencil tests. Although the prediction was not perfect and generalizability is limited, its simplicity and cost efficiency compensates for these limitations. Besides, response times are indispensable for the construction of test instruments in large-scale assessments and have to be gauged somehow. Our prediction formula offers a convenient way and might even outperform other methods such as expert ratings. However, users should carefully gauge if this prediction formula is suitable for their populations of persons and items. If predictions seem implausible, do not use them!

Acknowledgements

This work was supported by the Institute for Educational Quality Improvement at Humboldt-Universität zu Berlin, Berlin, Germany.

References

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Retrieved from http://lme4.r-forge.r-project.org/book/

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7) [Computer software]. Retrieved from http://CRAN.R-project.org/package=lme4

Bergstrom, B. A., Gershon, R. C., & Lunz, M. E. (1994, April). *Computer adaptive testing: Exploring examinee response time using hierarchical linear modeling*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution and display rate on computer-based test performance. *Applied Measurement in Education, 16*(3), 191–205.

Canty, A., & Ripley, B. (2014). boot: Bootstrap R (S-Plus) functions (Version 1.3-13) [Computer software]. Retrieved from http://CRAN.R-project.org/package=boot

Davison, M. L., Semmes, R., Huang, L., & Close, C. N. (2011). On the reliability and validity of a numerical reasoning speed dimension derived from response times collected in computerized testing. *Educational and Psychological Measurement, 72*(2), 245–263.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53.

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research & Perspective, 13*(3–4), 133–164.

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence, 39*(2–3), 108–119.

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015a). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement, 75*(6), 1021–1044.

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015b). Modeling booklet effects for non-equivalent group designs in large-scale assessment. *Educational and Psychological Measurement, 75*(4), 568–584.

Halkitis, P. N., Jones, J. P., & Pradhan, J. (1996, April). *Estimating testing time: The effects of item characteristics on response latency*. Paper presented at the Annual Meeting of the American Educational Research Association. New York, NY.

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. *International Journal of Environmental and Science Education, 4*(3), 275–288.

Klein Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*(1), 21–48.

Krämer, N., & Schäfer, J. (2014). parcor: Regularized estimation of partial correlation matrices (Version 0.2-6) [Computer software]. Retrieved from http://CRAN.R-project.org/package=parcor

Kremer, K., Fischer, H. E., Kauertz, A., Mayer, J., Sumfleth, E., & Walpuski, M. (2012). Assessment of standard-based learning outcomes in science education: Perspectives from the German project ESNAS. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible: Learning outcomes in science education* (pp. 201–218). Münster, Germany: Waxmann.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359–379.

Mayerl, J. (2005). Controlling the baseline speed of respondents: An empirical evaluation of data treatment methods of response latencies. In C. van Dijkum, J. Blasius, & B. van Hilton (Eds.), *Recent developments and applications in social research methodology. Proceedings of the sixth international conference on logic and methodology* (2nd ed.) [CD ROM]. Opladen, Germany: Barbara Budrich.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive-ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458.

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education, 8*(3), 545–563.

Overton, R. C., Taylor, L. R., Zickar, M. J., & Harms, H. J. (1996). The pen-based computer as an alternative platform for test administration. *Personnel Psychology, 49*(2), 455–464.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* [The IQB national assessment study 2012. Competencies in mathematics and the sciences at the end of secondary level]. Münster, Germany: Waxmann.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337–354.

R Core Team. (2014). R: A language and environment for statistical computing (Version 3.1.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models – Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Rutkowski, L., Gonzales, E., Davier, M. von, & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 75–95). Boca Raton, FL: CRC Press.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Mahwah, NJ: Lawrence Erlbaum Associates.

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment, 26*(4), 284–292.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*(5), 849–869.

Swanson, D. B., Case, S. M., Ripkey, D. R., Clauser, B. E., & Holtman, M. C. (2001). Relationships among item characteristics, examinee characteristics, and response times on USMLE Step 1. *Academic Medicine, 76*(Supplement 10), 114–116.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). New York, NY: Academic Press, Inc.

Van Lent, G. (2008). Important considerations in e-assessment. In F. Scheuermann, & A. Guimarães Pereira (Eds.), *Towards a research agenda on computer-based assessment. Challenges and needs for European educational measurement* (pp. 97–103). Luxembourg: European Commission. Retrieved from http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/907/1/reqno_jrc44526_report%20final%20version%5B2%5D.pdf

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*(5), 323–339.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5–24.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest 2.0 – Generalised item response modeling software*. Camberwell, England: ACER.